

## CLAIMS

What is claimed is:

- 1       1. A method of processing queries in a network, comprising the steps of:  
2               defining a plurality of virtual clusters from a plurality of servers; and  
3               routing a request to a virtual cluster of the plurality of virtual clusters based on  
4       predetermined criteria in order to allocate system resources.
  
- 1       2. The method of claim 1, further comprising the steps of:  
2               monitoring performance of the plurality of servers; and  
3               sending a report in response to workload at one of the plurality of servers  
4       exceeding a pre-determined threshold so that routing of further requests to the one of the  
5       plurality of servers is altered.
  
- 1       3. The method of claim 2, further comprising the step of removing the one of the plurality  
2       of servers from an associated virtual cluster and adding the one of the plurality of servers  
3       back into the associated virtual cluster in response to workload falling below the pre-  
4       determined threshold.
  
- 1       4. The method of claim 2, wherein the sending a report sends a report to a network dispatcher  
2       and the network dispatcher performs the routing.
  
- 1       5. The method of claim 1, further comprising the steps of:  
2               determining that one of the plurality of servers is overburdened based on  
3       statistics; and  
4               reducing workload to the one of the plurality of servers if the statistics are above a  
5       threshold.

- 1        6. The method of claim 5, wherein the reducing step includes at least one of removing the  
2        one of a plurality of servers from one of the plurality of virtual clusters and limiting further  
3        requests from being routed to the one of a plurality of servers.
- 1        7. The method of claim 6, wherein the reducing step includes reassigning the one of a  
2        plurality of servers to another one of the plurality of virtual clusters.
- 1        8. The method of claim 1, wherein at least one of the plurality of servers is assigned to more  
2        than one of the plurality of virtual clusters.
- 1        9. The method of claim 1, wherein the predetermined criteria includes at least one of  
2        requester identity, requested application, time of day, day of week, and performance  
3        statistics.
- 1        10. The method of claim 9, wherein the requester identity is an internet address.
- 1        11. The method of claim 9, wherein the performance statistics include at least one of central  
2        processing unit (CPU) performance statistics, memory statistics, connection counts,  
3        throughput statistics, and response time statistics.
- 1        12. The method of claim 1, wherein the routing step includes selecting one of the plurality of  
2        virtual clusters for routing based on at least one of a requester's identity and a requested  
3        application.
- 1        13. The method of claim 12, further including selecting one server from the one of the  
2        plurality of virtual clusters for routing based on statistics.
- 1        14. The method of claim 13, wherein the selecting is based on performance statistics.

1 15. The method of claim 1, wherein at least one of the plurality of servers is at least one of a  
2 lightweight directory access protocol (LDAP) server and a web application server.

1 16. The method of claim 1, wherein the routing uses rules based routing.

1 17. The method of claim 1, further comprising the steps of reassigning one of the plurality of  
2 servers from one of the plurality of virtual clusters to another one of the plurality of virtual  
3 clusters, wherein the one of the plurality of virtual clusters has a workload below a threshold  
4 and the another one of the plurality of virtual clusters has a workload above the pre-  
5 determined threshold.

1 18. A method for load balancing servers, comprising the steps of:  
2 allocating a plurality of servers among a plurality of virtual clusters;  
3 monitoring the plurality of virtual clusters for workload capacity; and  
4 reassigning at least one server from one of the plurality of virtual clusters to  
5 another of the plurality of virtual clusters based on workload capacity in order to reallocate  
6 system resources.

1 19. The method of claim 18, wherein the monitoring step includes determining when a  
2 workload capacity of the one of the plurality of virtual clusters has crossed a threshold based  
3 on statistics associated with the one of a plurality of virtual cluster's performance.

1 20. The method of claim 18, further comprising the step of identifying another of the  
2 plurality of virtual cluster having available workload capacity based on statistics associated  
3 with the virtual cluster's performance and transferring at least one of the plurality of servers  
4 to the another of the virtual cluster.

1 21. The method of claim 18, wherein the reassigning at least one server includes one of:

2 removing the server entirely from the one of a plurality of virtual cluster, and  
3 assigning the at least one server to both the one of a plurality of virtual clusters and the  
4 another of the plurality of virtual clusters.

1 22. The method of claim 18, further comprising routing a request to one of the plurality of  
2 virtual clusters based on one of the requestor's identity, the requested application, and rules.

1 23. The method of claim 22, further comprising selecting one server assigned to the one of  
2 the plurality of virtual clusters based on statistics for routing the request.

1 24. A computer program product comprising a computer usable medium having readable  
2 program code embodied in the medium, the computer program product includes at least one  
3 component to:

4 define a plurality of virtual clusters from a plurality of servers; and  
5 route a request to a virtual cluster of the plurality of virtual clusters based on  
6 predetermined criteria to allocate system resources.

1 25. The method of claim 24, wherein the at least one component:

2 monitors performance of the plurality of servers; and  
3 sends a report in response to workload at one of the plurality of servers exceeding  
4 a pre-determined threshold so that routing of further requests to the one of the plurality of  
5 servers is altered.

1 26. The system of claim 25, wherein the at least one component removes the one of the  
2 plurality of servers from an associated virtual cluster and adding the one of the plurality of  
3 servers back into the associated virtual cluster in response to workload falling below the pre-  
4 determined threshold.

1 27. The system of claim 24, wherein the at least one component sends a report to a network  
2 dispatcher and the network dispatcher performs the routing.

1 28. The system of claim 24, wherein the at least one component:  
2 determines that one of the plurality of servers is overburdened based on statistics;  
3 and  
4 reduces workload to the one of a plurality of servers if the statistics are above a  
5 threshold.

1 29. The system of claim 28, wherein the at least one component removes the one of a  
2 plurality of servers from one of the plurality of virtual clusters and limits further requests  
3 from being routed to the one of a plurality of servers.

1 30. The system of claim 29, wherein the at least one component reassigns the one of a  
2 plurality of servers to another one of the plurality of virtual clusters to reallocate the system  
3 resources.

1 31. The system of claim 24, wherein the at least one component assigns at least one of the  
2 plurality of servers to more than one of the plurality of virtual clusters.

1 32. The system of claim 24, wherein the predetermined criteria includes at least one of  
2 requester identity, requested application, time of day, day of week, performance statistics.

1 33. The system of claim 32, wherein the requester identity is a network address.

1 34. The system of claim 32, wherein the performance statistics include at least one of central  
2 processing unit (CPU) performance statistics, memory statistics, connection counts,  
3 throughput statistics, and response time statistics.

1 35. The system of claim 24, wherein the at least one component selects one of the plurality  
2 of virtual clusters for routing based on at least one of a requester's identity, composite  
3 statistics, and a requested application.

1 36. The system of claim 24, wherein the at least one component selects a non over-burdened  
2 server from the one of the plurality of virtual clusters to process information.

1 37. The system of claim 36, wherein the at least one component selects based on  
2 performance statistics.

1 38. The system of claim 24, wherein at least one of the plurality of servers is one of a  
2 lightweight directory access protocol (LDAP) server and a web application server.

1 39. The system of claim 24, wherein the at least one component uses rules based routing.

1 40. The system of claim 24, wherein the at least one component reassigns one of the plurality  
2 of servers from one of the plurality of virtual clusters to another one of the plurality of virtual  
3 clusters, wherein the another of the plurality of virtual clusters has a workload below a  
4 threshold and the one of the plurality of virtual clusters has a workload above the pre-  
5 determined threshold.